TEACHERS' ENGAGEMENT WITH A COMPETING MODELS INFORMAL INFERENCE TASK

Jennifer N. Lovett
Middle Tennessee State University
Jennifer.Lovett@mtsu.edu

Ryan Seth Jones
MTSU
Ryan.Jones@mtsu.edu

Matthew Duncan
MTSU
Matthew.Duncan@mtsu.edu

Informal inference is a critical practice for students to engage in if they are to understand formal statistical methods. However, during informal inference students often utilize complex ideas that many in-service teachers are not prepared for as they have not had the opportunity to think deeply about statistics and develop statistical knowledge for teaching (Groth, 2013). Research shows that engaging teachers in authentic inquiry of content supports the development of that content knowledge, and there is an urgent need to do so through professional developments (PDs). However, there is limited literature concerning PDs in statistics education, and a dearth of research focusing on teachers' engagement with informal inference tasks. This paper describes in detail how two teachers engaged with a seminal informal inference task during a PD, including their reasoning about variability and sample size when making inferences.

Keywords: Data Analysis and Statistics, Teacher Knowledge, Technology

Informal inference is a critically important practice for students to develop in middle and high school if they are to deeply understand more sophisticated inferential methods in later grades (Pfannkuch, 2011). Even if students do not go on to study formal inference this foundation will allow them to question claims presented to them in wider society. However, informal inference makes use of a complicated set of ideas, representations, and practices. What's more, research has shown that most teachers do not have a deep understanding of the foundational concepts related to statistical inference (Lovett & Lee, 2017; Franklin et al., 2015). These findings should not be viewed as shortcomings of teachers, but rather shortcomings of the educational experiences they received in their school mathematics and teacher preparation programs. Even though standards documents such as *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics, 2000) and *Common Core State Standards* (National Governors Association Center for Best Practice & Council of Chief State School Officers, 2010) have increased the emphasis on statistics, statistics has not been prioritized in K-12 mathematics classrooms and preservice mathematics teacher preparation programs. This has created an environment where teachers, even mathematics teachers, can complete their K-16 schooling without being supported to think deeply about statistics and develop the statistical knowledge for teaching (Groth, 2013) they need to teach the statistical standards in their curriculum. Research has shown that one effective way to increase teachers' content knowledge is to engage them in authentic inquiry that uses the concepts one desires teachers to develop (Borko, 2004). Thus there is an urgent need to support in-service teachers in deepening their statistical knowledge through professional development (PD) to support student learning in this area.

Even with this urgent need there has been a lack of research on PDs in statistics education. There are a few PDs that have researched teachers' statistical knowledge (Bargagliotti et al., 2014; Huey & Weber, 2018; Nieszporek, Biehler, & Griese, 2018; Peters, 2018; Peters, Watkins, & Bennett, 2014; Wassong, 2018) but none that we found report on teachers' approaches,

Otten, S., Candela, A. G., de Araujo, Z., Haines, C., & Munter, C. (2019). *Proceedings of the forty-first annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. St Louis, MO: University of Missouri.

understandings, and misunderstandings of specific informal inference tasks. These studies report pre/post knowledge gains or a retrospective analysis of the impact of the professional development on content (Huey & Weber, 2018; Nieszporek et al., 2018; Wassong, 2018), dilemmas teachers experienced regarding distributions or measures of center (Peters, 2018; Peters et al., 2014), or misunderstandings of sampling variability and regression that emerged as a result of several tasks (Bargagliotti et al., 2014). Thus the purpose of this paper is to share teachers' approaches and understandings on an informal inference task used during a professional development aimed to increase middle school teachers' statistical knowledge for teaching.

Background Literature and Framework

Informal Inference

A number of researchers have defined or described informal inference (e.g., Garfield & Ben-Zvi, 2008; Makar and Rubin, 2009; Rossman, 2008; Zieffler, Garfield, delMas, & Reading, 2008) and the common consensus is there are three key principles of informal inference: 1) claims, generalizations, predictions, parameter estimates, and conclusions, that extend beyond describing the given data; 2) the use of data as evidence for those claims; and 3) acknowledging uncertainty in their claims through the use of probabilistic language. Within informal inference Zieffler et al. (2008) identified three types of inference tasks that allow for the development of these key aspects of inferential reasoning: 1) estimate and draw a graph of a population based on a sample (referred to as *Samples and Populations*); 2) compare two or more samples of data to infer whether there is a real difference between the populations from which they were sampled (referred to as *Comparing Groups*); and 3) judge which of two competing models or statements in more likely to be true (referred to as *Competing Models*). Zieffler et al. (2008) identified three types of *competing models* tasks that have been used sample data to choose between two models or claims. In this paper we focus on one of these, a task used by Lee, Angotti, and Tarr (2010), the Schoolopoly task, which compares data generated by a probability device and the theoretical probability to determine if the probability device was "fair."

Schoolopoly task. The Schoolopoly task is a competing models task that asks students to investigate whether a die is fair or not based on a simulation (https://s3.amazonaws.com/fi-courses/tsdi/unit_3/Schoolopoly%20Task.pdf). There were a total of six dice companies to investigate and each pair of students were assigned one company. The students are asked to "investigate whether the die sent to you by the company is, in fact, fair. That is, are all six outcomes equally likely to occur?"

Variability

Research has shown that grounding statistical inquiry in a deep understanding of variability and supporting students to represent, describe, measure, and model variation in data can support them to understand statistical methods as well as the epistemic role they play in making claims in light of variability (e.g., delMas & Liu, 2005; Peters, 2011; Peters et al., 2014). In spite of the importance of helping students understand variability, research has also shown that variability is a topic that is challenging for people of all ages to grasp and the research on teachers' understanding of variation is sparse (Sánchez, daSilva, & Countinho, 2011). Random variation in everyday life is often perceived as having no structure, mere random differences. Because of this lay understanding of randomness, students and teachers alike often do not conceptualize random variation in ways that are consistent with statistical inquiry. For example, they rarely connect random variation with a mathematical structure that leverages probability to conceive of a

distribution of variable outcomes (Lehrer & Kim, 2009; Watson 2006). This challenge extends to the use of variation in order to make predictions of probability situations. Sánchez and Garcia (2008) asked six teachers to predict the number of times each number would occur if a dice was rolled 60 times. Five of the six teachers predicted an equal number of outcomes for each number on the dice; a sequence of 10, 10, 10, 10, 10, 10. These teachers were viewing variability deterministically, looking for a definite answer.

Sampling and Sample Size

When making decisions under uncertainty, most people tend to use judgmental heuristics that can cause errors in reasoning (Harradine, Batanero, & Rossman 2011). One example that applies to competing models tasks is that people tend to believe that even a small sample should reflect all of the characteristics of the population (Kahneman & Tversky, 1972). This misunderstanding can be extended to random samples. Even if a random sample is chosen in an appropriate way and of sufficient size, research has shown that students believe that the random sample is a replica of the population. This way of thinking does not take into consideration the variability across samples (Harradine et al., 2011; Saldanha & Thompson, 2002). Saldanha and Thompson (2002) found when high school students were asked to judge how representative a sample was in relation to a population parameter, students compared their sample to the population parameter and not "on how it might compare to a clustering of statistic's values" (p.265).

Professional Development Context

The task was situated within a year-long professional development program designed to blend online learning resources (*Teaching Statistics Through Data Investigations,* http://go.ncsu.edu/tsdi), an intensive summer workshop, and monthly professional learning community (PLC) meetings. The summer institute was designed to support the development of content knowledge and knowledge of student thinking using resources from a middle grades curriculum called Data Modeling (modelingdata.org). The year-long professional development framed statistical inquiry as the practice of generating knowledge using variable data, engaged teachers with statistical inquiry tasks, and supported them to analyze student artifacts and collectively plan instructional strategies for their classrooms.

The focus task in analysis occurred during day 3 of the five-day summer workshop. We began the first day of the workshop by engaging teachers in making an inference about a randomly drawn reward process to determine if the drawing was fair or biased in order to frame the activity of making model based inferences in the midst of variable systems. With this broad framing, we then engaged teachers with repeated measurements of a common object to engage with variability and distribution, and supported them to make judgements about student thinking using student artifacts from a similar activity. Teachers then explored students' approaches to measuring the center of a distribution, and teachers invented statistics to measure variability. During day 3 the teachers began by creating repeated samples of chance processes to explore sampling variability. The Schoolopoly task occurred after these activities, with the goal for teachers to use ideas about data, distribution, sampling variability, probability, and statistics to make their judgements about the dice companies.

Schoolopoly Task for Teachers

We modified Lee et al.'s (2010) original Schoolopoly task that was designed for middle school students to be used with middle school teachers during our PD (Figure 1). The first modification we made was in the choice of technology. Since Probability Explorer is not readily available to teachers, we used StatCrunch. The same six "companies" that Lee et al. (2010) used

are now built into StatCrunch as an app. Secondly, the teachers were asked to investigate all six dice companies instead of just the one company that middle school students were assigned. Finally, teachers were given a limited number of trials that they could assign to the six companies as they wanted to. In Lee et al.'s (2010) study, the middle school students were investigating the role of sample size for the first time and did not have any directions regarding the number of trials. However, since teachers were aware that a larger sample size would produce a more representative sample we chose to limit the number of trials they could run. Two additional parts of the task were added as the teachers were working through the task. They were unaware of these additional parts when they began working. Part B and C of the task allowed teachers additional trials to determine the one fair company. These decisions to modify the task for teachers were made so that teachers would have opportunities to compare empirical data from different populations, and so to provoke a need to make inferences with limited numbers of samples since students in the past often ran large numbers of trials to reduce uncertainty.

Part A: Your school is planning to create a board game modeled on the classic game of MonopolyTM. The game is to be called Schoolopoly and, like MonopolyTM, will be played with dice. Because many copies of the game expect to be sold, companies are competing for the contract to supply dice for Schoolopoly. Some companies have been accused of making poor quality dice and these are to be avoided since players must believe the dice they are using are actually "fair." Each company has provided a sample die for analysis and to investigate:

Luckytown Dice Company — Dice, Dice, Baby!
Dice R' Us — Pips and Dots
High Rollers, Inc. — Slice n' Dice

1. You have 300 trials to divide up between the 6 companies. Test the companies to determine if the company is fair or biased. There is at least one fair company.
2. In a google doc, record your evidence to support your decision if the company is fair or biased and your decision of how to divide up the 300 trials.

Part B: The PD facilitator offers each group 200 additional trials to identify the single unbiased company. Trials can be used to confirm or disconfirm their previous identifications of each company as biased or unbiased.

Part C: The PD facilitator offers each group 50 additional trials if they would like them.

Figure 1: Schoolopoly Task Modified from Lee et al. (2010).

Methodology

Given the importance of developing teachers' statistical knowledge for teaching and the lack of research on teachers' approaches to informal inference tasks, warrants study in the ways that teachers engage in an informal inference task during a PD. In this study we aim to answer the following research questions:

RQ 1 How do teachers engage with the Schoolopoly task?
RQ 2 Within the Schoolopoly task, how do teachers reason about sample size and variability?

To answer our research questions, we utilized a multiple case study design (Yin, 2009) to describe the ways teachers engaged with the Schoolopoly task and the ways they reasoned about

sample size and variability. Each case is represented by a pair of teachers working on the Schoolopoly task.

Participants

During the summer of 2017, 20 (2 = males, 18 = Females) middle school mathematics teachers participated in a week-long summer institute housed within year-long professional development project. Teachers were recruited from local Southeastern public schools representing seven different schools across both urban and suburban areas. Seven out of the twenty teachers were African-American and the remainder were White. As a whole, teaching experience ranged from less than one year to twenty plus years.

Data Collection and Analysis

Teachers each had access to a computer but worked in pairs to complete the Schoolopoly task. They collaborated in a google document to organize screenshots of the graphical representations created by StatCrunch as well as their identifications and justifications of each machine being fair or biased. We screen and audio recorded each pair during the task.

After the screen recordings were transcribed, each researcher individually wrote a narrative of the teachers working through the task. The narratives included portions of the transcript, justifications and graphical representations from the google document, and the researchers recording of their actions. Researchers compared their descriptions and collaboratively wrote a single description for each pair of teachers. Then we completed a cross-case comparison (Merriam, 1998) to identify themes in the ways teachers engaged with the task and how the reasoned about sample size and variability. Through this comparison three themes emerged; 1) The ways teachers articulated the two models under competition, 2) the ways teachers described and measured variability, and 3) the ways teachers accounted for sample size.

Results

Due to space limitations, we present one of the three cases of teachers, Myra and Kendall. This case was chosen because in many ways these teachers typified approaches of the other teachers in the PD, but also because of their efforts to formally quantify the variability to inform their decisions, which was less typical.

Kendall and Myra's Work on the Schoolopoly Task

Myra and Kendall began the task by dividing up their 300 roles evenly among the six companies before making any inferences about which companies were fair or biased. Through examination of the graphs Myra and Kendall immediately declared five of the six companies biased without much discussion. They made initial judgements about the status of each company by focusing on how similar the proportions are, explaining "If the dice were unbiased the results would be similar/same for each outcome."

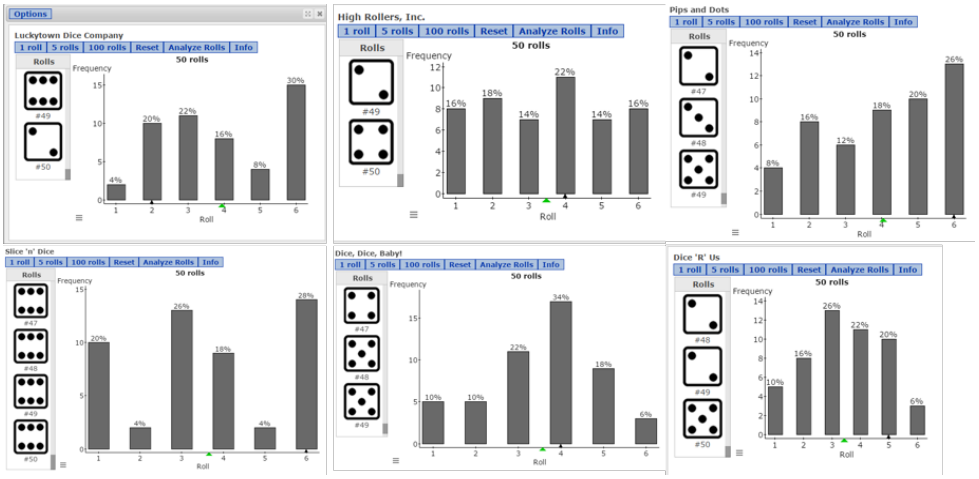


Figure 2: Myra and Kendall's Distributions After 50 Rolls

Myra and Kendall began looking for how similar outcomes were after the initial 50 rolls for each company. They then mathematized the idea of similarity by calculating the difference from the lowest and highest percentages for all six companies. For example, for Luckytown (Figure 2), Kendall and Myra wrote, "Biased. Luckytown has a range of results from 4% to 30%. This tells me that I am very unlikely to roll a 1 as compared to rolling a 6." This was their first method of quantifying the variability in the data. This approach led them to identify all companies but one as biased. The one company they declared unbiased after 50 rolls was High Rollers Inc (Figure 2). Kendall and Myra identified this company as fair since "the range of this data is 6%, making the results similar to each other (within 6% of each other) making this more likely to be the unbiased or fair option. They did not discuss sampling variability or mention what was an expected variation in the occurrences; only that six percent fell into the interval that they felt comfortable calling unbiased.

When Kendall and Myra were offered an additional 200 rolls and they first suggested to use 100 of their additional rolls on High Rollers Inc. to verify their results. Then they decided to consider other companies that may be fair even if they originally judged them as biased. To decide which companies to give additional rolls to they identified companies that had "no way" of being fair based on the first 50 rolls, referring to the difference in the highest and lowest percentages when making this decision. Kendall and Myra decided that a sample size of 50 was sufficient to make this decision as long as the difference was extreme.

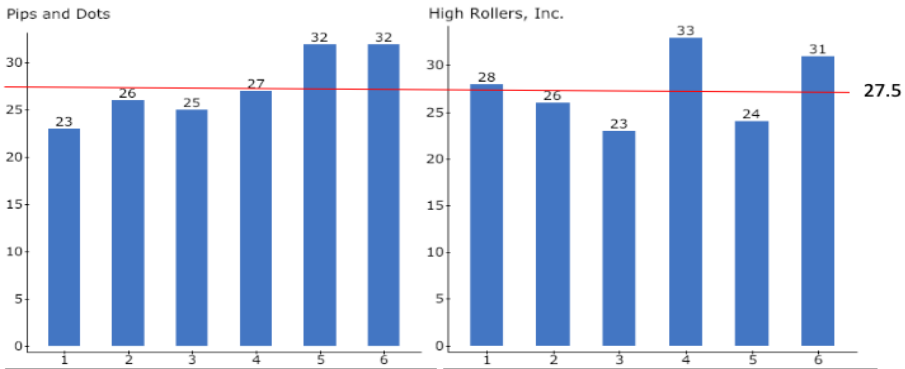
Kendall and Myra decided that the only companies that needed additional trials was Pips and Dots and High Rollers Inc. They give an additional 40 trials to Pips and Dots and High Rollers Inc. and continued with their method of comparing the difference in the percentages of the lowest to the highest occurrence. At this point they again compare all six companies, even though the two companies now have 90 trials and the other four still only have 50 trials. Kendall and Myra did not discuss how the sample size impacted the conclusions they were drawing. Instead continued to focus on companies that they "could for sure say absolutely not" unbiased,

> I feel like these percents are just so far apart like their range is four percent to 28 percent and there's two of them [referring to Slice and Dice and Dice R Us']. And this one we've got six percent to 34 percent [referring to Dice Dice Baby!]. I mean that's almost a 30 percent

difference on these ones. I mean this is 24 difference, this is almost 30. Like 28 and 11 to 22 that's only 11 difference. So are we, do we think Pips N Dots and High Rollers?

Myra and Kendall chose to only focus on High Rollers, Inc. and Pips and Dots. However, at this point in the task examining the differences in the percentages was not providing enough confirming evidence for Kendall and Myra to distinguish which one was fair, so they began looking at the expected values as a possible indication of bias. This was the first mention that the expected value of the dice, "So they should be in the 16 to 17 range so how far are they off from it? Because I feel like half of mine are pretty close to that and half of them are not." They continued with this line of thinking and continued adding trials to High Rollers Inc. and Pips and Dots until each company had a total of 140 trials. Myra and Kendall decided "to stick with our original company [High Rollers, Inc.] because it seemed fair with Part A and Part B where Pips and dots only seemed the most fair with Part B." However, they were still not convinced and wanted more trials.

Since Myra and Kendall were still unsure, the PD facilitator offered a final 50 trials. They decided to divide the trials evenly among High Rollers, Inc. and Pips and Dots for a total of 165 trials (Figure 3). Still trying to quantify the variability in the distributions, they determined the expected value for each number on the dice as 165/6 to be 27.5 trials. They then calculated "variability in the data by finding the average distance from 27.5." They referred to this measure as the M.A.D. For Pips and Dots this was 3 and for High Rollers, Inc. was 3.166. This Myra to comment that Pips & Dots was less variable and Kendall asked if that matters. Mayra said "it showed the outcomes were closer to the expectation if fair". They then agreed this measure was helpful and declared Pips and Dots a fair company.



Number	1	2	3	4	5	6
Frequency (f)	23	26	25	27	32	32
|f-27.5|	4.5	1.5	2.5	0.5	4.5	4.5
Average Absolute Deviations						3.0

Number	1	2	3	4	5	6
Frequency (f)	28	26	23	33	24	31
|f-27.5|	0.5	1.5	4.5	5.5	3.5	3.5
Average Absolute Deviations						3.17

Figure 3: Myra and Kendall's Measure of Variability After 165 Rolls

Articulating the competing models. During the task Kendall and Myra never explicitly articulated the two models under competition but as the task progressed there was a shift in their description of their null model. Originally, Kendall and Myra were comparing the distributions their data produced to a uniform probability model. They expected all outcomes for a single company to be "similar/same for each outcome." However, when using the difference in the highest and lowest outcome was not helping them move forward in identifying a single company, Kendall and Myra shifted their model. For the second half of the task, they described

their null model as the expected value either in terms of percentage or number of outcomes. This shift in their null model allowed Kendall and Myra to quantify the variability of each distribution as the dispersion from the expected value.

Describing and measuring variability. In the end, Kendall and Myra tried a total of three ways to describe the variability in the distributions: examining the difference in the highest and lowest percentage, examining the number of outcomes that fell into a middle region around the expected value, and calculated a measure by subtracting each frequency from the expected frequency of 27.5 and taking the average of those distances. These approaches increased in their sophistication when they had to identify a single company as fair. Their first two attempts at quantifying the variability were subjective by comparing the differences in the highest and lowest percentages and the number of outcomes that fell into the middle region. For both of these they subjectively determined an acceptable range in percentages or what constituted the middle region around the expected value. Their final approach, subtracting each frequency from the expected frequency of 27.5 and taking the average of those distances, provide Kendall and Myra a numerical measure to compare the variability in each companies' distributions.

Accounting for sample size. Sample size played a crucial role in the way Kendall and Myra engaged with the task and how they discussed variability. The way in which Kendall and Myra accounted for the sample size at the beginning of the task, evenly distributing the samples, was similar to the way the majority of the teachers engaged with the task. Throughout the task Kendall and Myra appeared to be ok with making a conclusion that a company was biased from a sample of 50 but wanted a larger sample to be able to determine if a company was fair. This was seen when they were comparing Pips and Dots and High Rollers Inc., that each had 90 trials, to the other four companies that only had their original 50 trials. Kendall and Myra continued to classify the other companies as biased with a smaller number of trials. Their struggle with accounting for sample size in their conclusions was also shown when they decided "to stick with our original company [High Rollers, Inc.] because it seemed fair with Part A and Part B where Pips and Dots only seemed the most fair with Part B." Even though Kendall and Myra had an understanding that a larger sample size is better they did not seem to understand that larger samples would produce less sampling variability. Overall, when Kendall and Myra were considering sample size their logic seemed to focus on rolling more trials to find evidence to support the fairness, and not how much variation is expected when a company is fair.

Discussion and Conclusion

Due to the infancy of research in statistics education, many tasks that statistics educators use with teachers in a PD have not been researched with that population. We chose to incorporate the Schoolopoly task into our summer PD to engage teachers in an informal inference task that required them to use ideas about data, distribution, sampling variability, probability, and statistics to make their judgements about the competing models. Thus, we examined the ways in which teachers engaged with the task and reasoned about sample size and variability. In doing so we identified three themes emerged; 1) The ways teachers articulated the two models under competition, 2) the ways teachers described and measured variability, and 3) the ways teachers accounted for sample size.

From our results, there are several changes we would make to our modified version of Schoolopoly task for the future. It is evident that Myra and Kendall expect that a random sample of sufficient size should replicate the population characteristics (Harradine et al., 2011). Therefore, even though teachers know that large samples are more representative of the

population they need more opportunities to engage in instructional activities to draw many samples to understand the variability among samples from a population (delMas, Garfield, & Chance, 1999; Saldanha & Thompson, 2002). In the end, we would still limit the number of trials the teachers could use to explore the companies but instead of adding additional trials to their distributions, have teachers generate additional samples of 50 trials and then add to their new samples. This would allow teachers to still examine all of the companies and compare multiple samples of 50 for the same company before determining which companies are fair, providing them the opportunity to estimate the underlying probability distribution of each company, similar to Lee et al.'s (2010) study. Then teachers could continue with the task, examining larger samples and comparing those samples to make a conclusion on which company they believe is fair.

References

Bargagliotti, A., Anderson, C., Casey, S., Everson, M., Franklin, C., Gould, R., . . . Watkins, A. (2014). Project-SET materials for the teaching and learning of sampling variability and regression. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July 2014) Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3). Retrieved from http://www.amstat.org.prox.lib.ncsu.edu/PUBLICATIONS/JSE/secure/v7n3/delmas.cfm

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55-82.

Franklin, C., Bargagliotti, A. E., Case, C. A., Kader, G. D., Schaeffer, R. L., & Spangler, D. A. (2015). *The statistical education of teachers*: American Statistical Association.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York: Springer.

Groth, R. E. (2013). A day in the life of a statistical knowledge for teaching course. *Teaching Statistics, 35*(1), 37-42.

Harradine A., Batanero C., & Rossman A. (2011) Students' and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics: Challenges for Teaching and Teacher Education* (pp. 235-246). Dordrecht, The Netherlands: Springer.

Huey, M. E., & Weber. W. (2018). Strategies employed by secondary mathematics teachers on inferential reasoning tasks. In M. A. Sorto, A. White, & L. Guyout (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 2018), Kyoto, Japan.* Voorburg: The Netherlands: International Statistical Institute. Retrieved from: http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_C201.pdf.

Kahneman, D., & Tversky, A. (1972) Subjective probability: A judgement of representativeness. *Cognitive Psychology 3*, 430–454.

Lee, H. S., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: Students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal, 9*(1), 68 - 96.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal, 21*(2), 116-133.

Lovett, J. N., & Lee, H. S. (2017). New standards require teaching more statistics: Are preservice secondary mathematics teachers ready? *Journal of Teacher Education, 68*(3), 299-311.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practice & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington D.C.: Author.

Nieszporek, R., Biehler, R., & Griese, B. (2018). Developments of teachers' knowledge facets in teaching statistics with digital tools measured with retrospective self-assessment. In M. A. Sorto, A. White, & L. Guyout (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 2018), Kyoto, Japan.* Voorburg: The Netherlands: International Statistical Institute. Retrieved from:http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_C103.pdf.

Peters, S. A. (2011). Robust understandings of variation. *Statistics Education Research Journal, 10*(1), 52-88.

Peters, S. A. (2018). Professional development to transform middle and high school teachers' understandings about distribution. In M. A. Sorto, A. White, & L. Guyout (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 2018), Kyoto, Japan.* Voorburg: The Netherlands: International Statistical Institute. Retrieved from: http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_4G3.pdf.

Peters, S. A., Watkins, J. D., & Bennett, V. M. (2014). Middle and high school teachers' transformative learning of center. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July 2014) Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. Mathematical Thinking and Learning, 13(1–2), 27–46.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5-19.

Sánchez, E., da Silva, C. B., & Coutinho, C. (2011). Teachers' understanding of variation. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education* (pp. 211-221). The Netherlands: Springer.

Sánchez, E., & Garcia, J. (2008). Acquisition of notions of statistical variation by in-service teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference*. Monterrey, Mexico: International Commission on Mathematical Instructions and International Association for Statistical Education.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257 - 270.

Tarr, J. E., Lee, H. S., & Rider, R. (2006). When data and chance collide: Drawing inferences from empirical data. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance: 2006 yearbook of the NCTM* (pp. 139-149). Reston, VA: National Council of Teachers of Mathematics.

Wassong, T. (2018). What kind of content remains in memory after a continuous professional development for statistics? Results of an interview study. In M. A. Sorto, A. White, & L. Guyout (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 2018), Kyoto, Japan.* Voorburg: The Netherlands: International Statistical Institute. Retrieved from:http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_4G2.pdf.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals.* Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: Sage.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58.